

On convergence rates of robust adaptive game theoretic learning algorithms

Zhisheng Hu ^a, Minghui Zhu ^a, Ping Chen ^b, Peng Liu ^b

^a*School of Electrical Engineering and Computer Science, Pennsylvania State University, 201 Old Main, University Park, PA, 16802, USA*

^b*College of Information Sciences and Technology, Pennsylvania State University, 201 Old Main, University Park, PA, 16802, USA*

Abstract

This paper investigates a class of multi-player discrete games where each player aims to maximize its own utility function. Two particular challenges are considered. Firstly, each player is unaware of the structure of its utility function and the actions of other players, but is able to access the corresponding utility value given an action profile. Second, utility values are subject to delays and errors. We propose a robust adaptive learning algorithm which converges in probability to the set of action profiles which have maximal stochastic potential. Furthermore, the convergence rate of the proposed algorithm is quantified. When the interactions of the players consist of a weakly acyclic game, the convergence to the set of pure Nash equilibria is guaranteed. A set of numerical simulations are conducted to validate the algorithm performance.

Key words: Distributed control; Game theory; Learning in games

1 Introduction

Game theory provides a mathematically rigorous framework for multiple players to reason about each other. In recent years, game theoretic learning has been increasingly used to control large-scale networked systems due to its inherent distributed nature. In particular, the network-wide objective of interest is encoded as a game whose Nash equilibria correspond to desired network-wide configurations. Numerical algorithms are then synthesized for the players to identify Nash equilibria via repeated interactions. Multi-player games can be categorized into discrete games and continuous games. In a discrete (resp. continuous) game, each player has a finite (resp. an infinite) number of action candidates. As for discrete games, learning algorithms include best-response dynamics, better-response dynamics, fictitious play, regret matching, logit-based dynamics and replicator dynamics. Please refer to (Basar and Olsder, 1999; Fudenberg and Levine, 1998; Sandholm, 2010; Young, 2001) for detailed discussion. As an important class of

continuous games, generalized Nash games were first formulated in Arrow and Debreu (1954), and see survey paper Facchinei and Kanzow (2007) for a comprehensive exposition. A number of algorithms have been proposed to compute generalized Nash equilibria, including, to name a few, ODE-based methods Rosen (1965), nonlinear Gauss-Seidel-type approaches Pang et al. (2008), iterative primal-dual Tikhonov schemes Yin et al. (2011), and best-response dynamics Palomar and Eldar (2010). Game theory and its learning have found many applications; e.g., traffic routing in Internet Altman et al. (2002), urban transportation Rouboutsos and Kapros (2008), mobile robot coordination (Arslan et al., 2007; Hatanaka et al., 2016) and power markets (Wang et al., 2012; Zhu, 2014).

In many applications, players can only access limited information about the game of interest. For example, each player may not know the structure of its own utility function. Additionally, during repeated interactions, each player may not be aware of the actions of other players. These informational constraints motivate recent study on adaptive learning algorithms where the players adjust their actions only based on their own previous actions and utility measurements. The papers (Marden et al., 2009; Zhu and Martinez, 2013; Hatanaka et al., 2016) study discrete games, and their approaches are based

* This work was partially supported by ARO W911NF-13-1-0421 (MURI) and NSA H98230-15-1-0289.

Email addresses: `zsh128@psu.edu` (Zhisheng Hu), `muz16@psu.edu` (Minghui Zhu), `pzc10@ist.psu.edu` (Ping Chen), `pliu@ist.psu.edu` (Peng Liu).

on stochastic stability Foster and Young (1990) and resistance tree theory Young (1993). The papers (Frihauf et al., 2012; Liu and Krstic, 2011; Stankovic et al., 2012) investigate continuous games. The papers (Frihauf et al., 2012; Liu and Krstic, 2011; Stankovic et al., 2012) employ extremum seeking and the paper Zhu and Frazzoli (2016) uses finite-difference approximations to estimate unknown gradients. However, none of these papers quantifies the convergence rates of the proposed algorithms. In addition, most of the existing work do not consider the errors and delays of utility measurements. One exception is Zhu and Frazzoli (2016) where delays are studied for generalized Nash games.

Contribution: In this paper, we study a class of multi-player discrete games. The particular challenges we consider include (i) each player is unaware of the actions taken by the others and its own utility function; (ii) its received utilities are subject to finite delays and errors. We propose a robust adaptive learning algorithm where, at each iteration, each player, on one hand, exploits successful actions in recent history via comparing received utility values, and on the other hand, randomly explores any feasible action with a certain exploration probability. The history length is determined by the maximum delay. The algorithm is proven to be convergent in probability to the set of action profiles with maximum stochastic potential. Furthermore, the convergence rate is formally quantified in terms of the utility delays, errors and parameters in our algorithm. When the interactions of the players consist of a weakly acyclic game, the convergence to the set of pure Nash equilibria is guaranteed. Two case studies are conducted to evaluate the algorithm performance. A preliminary version of this paper was published in Zhu et al. (2014) where convergence rates, utility errors and delays were not discussed. Further, Zhu et al. (2014) focuses on the application on adaptive cyber defense, and this paper focuses on theory of learning in games. The analysis of two papers is significantly different.

2 Problem formulation and learning algorithm

In this section, we introduce a class of multi-player games where the information each player accesses is limited. Then, we present a learning algorithm under which the action profiles of the players converge to the set of action profiles which have maximum stochastic potential.

2.1 Game formulation

The system model in Figure 1 is proposed to characterize the interactions of N players in a non-cooperative game. Each component in the figure will be discussed in the following paragraphs.

Players. We consider N players $\mathcal{I} = \{1, \dots, N\}$ and each player has a finite set of actions. Let \mathcal{A}_i denote the

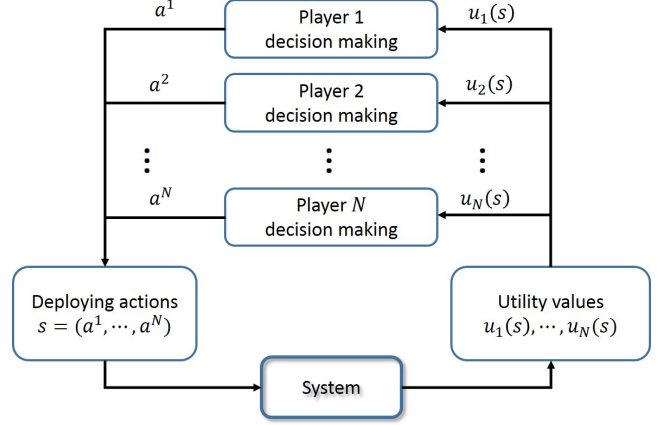


Fig. 1. Game model

action set of player i and $a^i \in \mathcal{A}_i$ denote an action of player i . Denote $\mathcal{S} \triangleq \mathcal{A}_1 \times \dots \times \mathcal{A}_N$ as the Cartesian product of the action sets, where $s \triangleq (a^1, \dots, a^N) \in \mathcal{S}$ is denoted as an action profile of the players.

Utility. Under the influence of an action profile, the system generates a utility value for each player. The utility function for player $i \in \mathcal{I}$ is defined as $u_i : \mathcal{S} \rightarrow \mathbb{R}$. At the end of iteration t , the utility value $u_i(t) = u_i(s(t))$ is measured and sent to player i . The utility measurements could be subject to errors and the utility transmissions could be subject to delays. So, for iteration t , the utility values received by player i are denoted by the vector $\bar{u}_i(t) = [\tilde{u}_i(t - \mathcal{T}_i^1) \ \tilde{u}_i(t - \mathcal{T}_i^2) \ \dots \ \tilde{u}_i(t - \mathcal{T}_i^{h_i(t)})]^T$, where $\mathcal{T}_i^1, \dots, \mathcal{T}_i^{h_i(t)}$ are the nonnegative delays, $\tilde{u}_i(t - \mathcal{T}_i^\ell) \triangleq u_i(t - \mathcal{T}_i^\ell) + e_o^i(t - \mathcal{T}_i^\ell)$ is the received utility value containing error $e_o^i(t - \mathcal{T}_i^\ell)$ and $h_i(t)$ is the nonnegative number of the utility values received by the player at iteration t . If player i receives nothing, then $h_i(t) = 0$ and $\bar{u}_i(t)$ is empty. Notice that each utility value is only received once. We assume that the utility delays for all players are uniformly upper bounded by \mathcal{T} .

Information constraint. In this paper, we consider the case that each player does not know the other players' action sets or their deployed actions. Besides, each player is unaware of the structure of its own or the others' utility functions. At iteration t , each player only knows its deployed actions up to iteration t and its received utility values. That is, at iteration t , player i is only aware of $a^i(1), \dots, a^i(t-1), \bar{u}_i(1), \dots, \bar{u}_i(t-1)$.

The above information constraint has been studied in several recent papers. For example, the authors in (Zhu and Martinez, 2013; Stankovic et al., 2012; Hatanaka et al., 2016) investigate coverage optimization problems for mobile sensor networks where mobile sensors are unaware of environmental distribution functions. The authors in Marden et al. (2013) study the problem of optimizing energy production in wind farms where each

turbine knows neither the functional form of the power generated by the wind farm nor the choices of other turbines. The authors in (Frihauf et al., 2012; Zhu and Frazzoli, 2016) consider convex games where each player cannot access its game components.

2.2 Problem statement

Under the above informational constraint, we aim to synthesize learning algorithms under which the action profiles of the players converge to the set of action profiles with maximum stochastic potential. In contrast to existing work, our algorithm is robust to utility delays and errors. Further, we will quantify the convergence rate of the proposed algorithm in contrast to asymptotic convergence in existing work.

2.3 Learning algorithm

Inspired by Zhu and Martinez (2013), we propose a learning algorithm where each player updates its actions only based on its previous actions and the received utility values. On the one hand, each player chooses the most successful action in recent history. It represents the exploitation phase. However, the exploitation is not sufficient to guarantee that the player can choose the most successful action within its whole action set. So on the other hand, the player uniformly chooses one action from its action set. It represents the exploration phase. The specific update rule is stated in Algorithm 1. Note that all the players can only exploit the actions after they receive their corresponding utility values. And the players surely can receive the utility values measured $\mathcal{T} + 1$ iterations ago. At the first $2\mathcal{T} + 1$ iterations, each player uniformly chooses one action from its action set (Line 3). Starting from iteration $2\mathcal{T} + 2$, with probability $1 - \tilde{\epsilon}_i(t)$, player i compares the two actions chosen $\mathcal{T} + 1$ and $2\mathcal{T} + 2$ iterations ago, and then chooses the one with a higher utility value as current action (Line 8-13). This represents the exploitation where player i reinforces its previous successful actions. With probability $\tilde{\epsilon}_i(t)$, player i uniformly explores the set \mathcal{A}_i (Line 14). This represents the exploration where player i tries all possible actions. The notations in Algorithm 1 are listed as follows:

- $\epsilon_i(t)$: the intended exploration rate of player i and it is considered as the control parameter of player i ;
- $e_c^i(t)$: the control error at iteration t for player i ;
- $\tilde{\epsilon}_i(t) \triangleq \epsilon_i(t) + e_c^i(t)$: the corrupted exploration rate of player i at iteration t ;
- $\text{sample}(\mathcal{A}_i)$ represents uniformly choosing one element from set \mathcal{A}_i .

3 Analysis

In this section, we will summarize the analytical results of our learning algorithm.

Algorithm 1. Robust adaptive learning algorithm

```

1: while  $0 \leq t \leq 2\mathcal{T} + 1$  do
2:   for  $i \in \mathcal{I}$  do
3:      $a^i(t) \leftarrow \text{sample}(\mathcal{A}_i)$ ;
4:   end for
5: end while
6: while  $t \geq 2\mathcal{T} + 2$  do
7:   for  $i \in \mathcal{I}$  do
8:     With prob.  $(1 - \tilde{\epsilon}_i(t))$ ,
9:     if  $\tilde{u}_i(t - \mathcal{T} - 1) \geq \tilde{u}_i(t - 2\mathcal{T} - 2)$  then
10:       $a^i(t) = a^i(t - \mathcal{T} - 1)$ ;
11:     else
12:       $a^i(t) = a^i(t - 2\mathcal{T} - 2)$ ;
13:     end if
14:     With prob.  $\tilde{\epsilon}_i(t)$ ,  $a^i(t) \leftarrow \text{sample}(\mathcal{A}_i)$ ;
15:   end for
16: end while

```

3.1 Notations and assumptions

Before the analysis of the algorithm, we first introduce some notations and assumptions for the rest of the paper. Denote $\epsilon(t) \triangleq (\epsilon_1(t), \dots, \epsilon_N(t))$ and $\epsilon_v(t) \triangleq \max\{\tilde{\epsilon}_1(t), \dots, \tilde{\epsilon}_N(t)\}$. We assume the exploration rates satisfy:

- Assumption 1** (1) For all $i \in \mathcal{I}$, $\epsilon_i(t)$ is non-negative, strictly decreasing, and $\lim_{t \rightarrow \infty} \epsilon_i(t) = \epsilon_i^* = 0$;
- (2) For any $i \neq j$, there exists a pair of $\gamma_{ij}, \gamma_{ji} > 0$ such that $\frac{\epsilon_i(t)}{\epsilon_j(t)} = \gamma_{ij}$ and $\frac{\epsilon_j(t)}{\epsilon_i(t)} = \gamma_{ji}$;
- (3) The sequence $\{\prod_{i=1}^N \epsilon_i(t)\}$ is not summable;

Assumption 1 indicates that as the game goes on, all the players' exploration rates decrease in the same order to 0 and the exploration rates do not decrease too fast. The slowly decreasing exploration rates ensure that the action profile converges to desired configurations asymptotically not matter where the game starts.

The vector norm is L^1 -norm, and the matrix norm is $\|P\| \triangleq \max_i \sum_k |p_{ik}|$.

Markov chains induced by Algorithm 1. Denote by $\mathcal{Z} \triangleq \mathcal{S} \times \mathcal{S}$ the state space, where each state $z(t) \triangleq (s(t), s(t + (\mathcal{T} + 1)))$ consists of the action profile at iteration t and the action profile $\mathcal{T} + 1$ iterations later. We will study the convergence properties of the action profiles by analyzing the transient behavior of the sequence $\{z(t)\}$. The sequence of iterations $\{t\} \subset \mathbb{N}^0$ is partitioned into $\mathcal{T} + 1$ sub-sequences, where each sub-sequence is denoted as $\{t_q\} \subset \mathbb{N}_q^0$ where $\mathbb{N}_q^0 \triangleq \{q + j(\mathcal{T} + 1) | j \in \mathbb{N}^0\}$ and $q \in \{0, 1, \dots, \mathcal{T}\}$. The relation between the sub-sequences t_q and the sequence t is shown in Figure 2, where the time unit in each sub-sequence is $\mathcal{T} + 1$ iterations. Then the sequence $\{z(t)\}$ can be partitioned

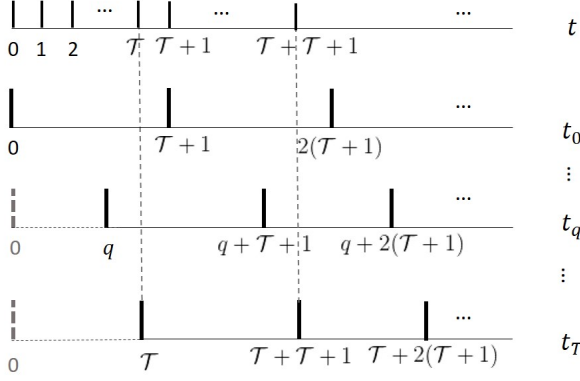


Fig. 2. Partition of time horizon

into $\mathcal{T} + 1$ sub-sequences; e.g., the sub-sequence with respect to $\{t_q\}$ is $\{z(q), z(q + \mathcal{T} + 1), \dots\}$. By the definition of z , each sub-sequence actually forms a time-inhomogeneous Markov chain. Denote the $\mathcal{T} + 1$ Markov chains as $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_{\mathcal{T}}$ and define $P_q^{\epsilon(t_q)}$ as the transition matrix of Markov chain \mathcal{M}_q at iteration $t = t_q$, where each entry $P_q^{\epsilon(t_q)}(z', z)$ represents the transition probability from state z' to z . Besides, denote by $\pi(t_q)$ the distribution on \mathcal{Z} at iteration t_q .

z -tree of Markov chain \mathcal{M}_q at iteration t_q . For $t = t_q$, given any two distinct states z' and z of Markov chain \mathcal{M}_q , consider all paths starting from z' and ending at z . Denote by $p_{z'z}^q$ the largest probability among all possible paths from z' to z . A path might contain intermediate states z_1, \dots, z_k ($k = 0$ means there is no intermediate state) between z' and z . So $p_{z'z}^q$ is the product of $P_q^{\epsilon(t_q)}(z', z_1), P_q^{\epsilon(t_q)}(z_1, z_2), \dots, P_q^{\epsilon(t_q)}(z_k, z)$. We define graph $\mathcal{G}(t_q)$ where each vertex of $\mathcal{G}(t_q)$ is a state z of Markov chain \mathcal{M}_q and the probability on edge (z', z) is $p_{z'z}^q$. A z -tree on $\mathcal{G}(t_q)$ is a spanning tree such that from every vertex $z' \neq z$, there is a unique path from z' to z . Denote by $G_q(z)$ the set of all z -trees on $\mathcal{G}(t_q)$. The total probability of a z -tree is the product of the probabilities of its edges. The *stochastic potential* of the state z is the largest total probability among all z -trees in $G_q(z)$. Let $\Lambda(\epsilon(t))$ be the set of states which have maximum stochastic potential at particular iteration t . And denote by $\Lambda(\epsilon^*)$ the set of states which have maximum stochastic potential as $\epsilon(t) \rightarrow 0$.

Remark 1 The above notions are inspired by the resistance trees theory Young (1993). However, the above notions are defined for any $\epsilon(t) \in (0, 1)$ instead of $\epsilon \rightarrow 0$ in the resistance trees theory. This allows us to characterize the transient performance of our algorithm. \square

For all $i \in \mathcal{I}$, let $u_{\wedge}^i \triangleq \frac{1}{2} \min_{s, s' \in \mathcal{S}, u_i(s) \neq u_i(s')} |u_i(s) - u_i(s')|$. Assume the utility errors satisfy:

Assumption 2 There exists $t_e > 2\mathcal{T} + 2$, such that for

all $i \in \mathcal{I}$, $|e_o^i(t)| < u_{\wedge}^i$ for all $t \geq t_e$.

Assumption 2 implies that as time goes on, the utility errors will be small enough so that the players will not confuse any two action profiles with different utility values.

Denote by $e_c^{\vee}(t) \triangleq \max\{e_c^1(t), \dots, e_c^N(t)\}$ the maximum control error at iteration t . Assume the maximum control errors satisfy:

Assumption 3 The sequence of maximum control errors $e_c^{\vee}(t)$ is absolutely summable; i.e., $\sum_{t=1}^{\infty} |e_c^{\vee}(t)| < \infty$.

Assumption 3 requires that the control errors decrease to zero fast enough.

3.2 Main result

The following theorem is the main analytical result of this paper. In particular, it shows that the state $z(t)$ asymptotically converges to the set of states with maximum stochastic potential in probability. The convergence is proved by the convergence of the sequence of state distribution $\{\pi(t)\}$ to the limiting distribution $\pi^* \triangleq \lim_{t \rightarrow \infty} \pi(t)$ whose support is contained in $\Lambda(\epsilon^*)$.

Moreover, the convergence rate is quantified in terms of the utility delays, errors and parameters in our algorithm. In particular, we define the summation of the distances between the state distributions $\pi(t)$ and the limiting distribution π^* during a period containing

$$\mathcal{T} + 1 \text{ iterations; i.e., } D(j) \triangleq \sum_{q=0}^{\mathcal{T}} \|\pi(q + j(\mathcal{T} + 1)) - \pi^*\|,$$

where $j \in \mathbb{N}^0$. The formal proof of Theorem 1 will be given in Section 4.

Theorem 1 Let Assumptions 1 - 3 hold. Consider the sequence of states $\{z(t)\}$ induced by Algorithm 1. Then the following properties hold:

- (P1) $\pi(t)$ converges to limiting distribution π^* ;
- (P2) the support of π^* is contained in $\Lambda(\epsilon^*)$ and $z(t)$ converges to $\Lambda(\epsilon^*)$ in probability;
- (P3) there exists some $t^{\vee} \in \mathbb{N}^0$ such that for any $t^* > t^{\vee}$ and $(j-1)(\mathcal{T} + 1) \geq t^*$, the convergence rate is quantified as follows:

$$\begin{aligned} D(j) \leq (\mathcal{T} + 1) & \left(C_e \exp\left(-\prod_{i=1}^N |\mathcal{A}_i|\right) \sum_{\tau=t^*}^{(j-1)(\mathcal{T}+1)} \prod_{i=1}^N \epsilon_i(\tau) \right. \\ & + 4C_{\vee} \epsilon_{\vee}(t^*) + C_{\vee} \epsilon_{\vee}(j(\mathcal{T} + 1)) \\ & \left. + 2^N (2^N - 1) \prod_{i=1}^N |\mathcal{A}_i| \sum_{\tau=t^*}^{\mathcal{T}+j(\mathcal{T}+1)} |e_c^{\vee}(\tau)| \right), \end{aligned} \quad (1)$$

where $C_v > 0$ and $C_e > 0$ and the summation $\sum_{(j-1)(\mathcal{T}+1)}^N$ is to sum up the entries every $\mathcal{T} + 1$ iterations; e.g.,

$$\sum_{\tau=t^*}^N \prod_{i=1}^N \epsilon_i(\tau) \triangleq \prod_{i=1}^N \epsilon_i(t^*) + \prod_{i=1}^N \epsilon_i(t^* + \mathcal{T} + 1) + \dots + \prod_{i=1}^N \epsilon_i((j-1)(\mathcal{T} + 1)).$$

3.3 Discussion

Explicit convergence rate. If the exploration rates, utility delays and control errors are given, we can explicitly quantify how fast the algorithm will reach the set $\Lambda(\epsilon^*)$. Assume the exploration rate for player i is $(|\mathcal{A}_i|)^{-1} t^{-\frac{1}{N}}$, and the control errors for all players are $\left(2^N(2^N - 1)t^2 \prod_{i=1}^N |\mathcal{A}_i|\right)^{-1}$ and the utility delay upper bound \mathcal{T} is 1. Given any small $\delta > 0$ such that $\left(\frac{4(4C_v + C_v + 1)}{\delta}\right)^N \geq t^v$, then $D(j) < \delta$ for all $t^* > \left(\frac{4(4C_v + C_v + 1)}{\delta}\right)^N$ and $j \geq \frac{8C_e^2 t^*}{\delta^2} + 1$. The verification is given as follows. By Theorem 1, we have

$$\begin{aligned} D(j) &\leq 2 \left(C_e \exp \left(- \sum_{\tau=t^*}^{2(j-1)} \frac{1}{\tau} \right) + 4C_v \epsilon_v(t^*) + C_v \epsilon_v(2j) \right. \\ &\quad \left. + \sum_{\tau=t^*}^{1+2j} \frac{1}{\tau^2} \right) \\ &\leq 2 \left(C_e \exp \left(- \frac{1}{2} \int_{t^*}^{2(j-1)} \frac{1}{x} dx \right) + 4C_v (t^*)^{-\frac{1}{N}} |\mathcal{A}_i|^{-1} \right. \\ &\quad \left. + C_v (2j)^{-\frac{1}{N}} |\mathcal{A}_i|^{-1} + \frac{1}{t^*} + \frac{1}{2} \left(\frac{1}{t^*} - \frac{1}{1+2j} \right) \right) \\ &< 2 \left(C_e \sqrt{\frac{t^*}{2(j-1)}} + (4C_v + C_v)(t^*)^{-\frac{1}{N}} |\mathcal{A}_i|^{-1} + \frac{2}{t^*} \right) \\ &< 2C_e \sqrt{\frac{t^*}{2(j-1)}} + 2(4C_v + C_v + 1)(t^*)^{-\frac{1}{N}}. \end{aligned}$$

Then we have $2(4C_v + C_v + 1)(t^*)^{-\frac{1}{N}} \leq \frac{\delta}{2}$ for all $t^* \geq \left(\frac{4(4C_v + C_v + 1)}{\delta}\right)^N$ and $2C_e \sqrt{\frac{t^*}{2(j-1)}} \leq \frac{\delta}{2}$ for all $j \geq \frac{8C_e^2 t^*}{\delta^2} + 1$.

Weakly acyclic game. In Section 3.2, we do not specify whether a pure Nash equilibrium exists. In this section, we study a special case where the interactions of the players consist of a weakly acyclic game. It is known that any weakly acyclic game has at least one pure Nash equilibrium (Fabrikant et al., 2010; Milchtaich, 1996; Young, 1993).

Definition 1 (Pure Nash equilibrium) An action profile $s^* \triangleq (a_*^1, \dots, a_*^i, \dots, a_*^N)$ is a pure Nash equilibrium

if no player can benefit from unilateral deviations. That is, $\forall i \in \mathcal{I}, \forall a_*^i \in \mathcal{A}_i, u_i(a_*^1, \dots, a_*^i, \dots, a_*^N) \geq u_i(a_*^1, \dots, a_*^i, \dots, a_*^N)$.

Definition 2 (Weakly acyclic game) A game is called to be weakly acyclic if it satisfies that from every action profile s , there exists a finite best-response improvement path leading from that action profile to a pure Nash equilibrium.

Remark 2 In Section 5, we will study a real-world cyber security scenario which is a weakly acyclic game. The verification will be based on a simple criterion given in Takahashi and Yamamori (2002) instead of using Definition 2. \square

Denote the set of pure Nash equilibria of the game Γ as $\mathcal{E}(\Gamma)$ and $\text{diag}(\mathcal{E}(\Gamma) \times \mathcal{E}(\Gamma)) \triangleq \{(s, s) \in \mathcal{Z} | s \in \mathcal{E}(\Gamma)\}$. It is proved in the Claims 2-4 in the Proposition 4.3 of Zhu and Martinez (2013) that $\Lambda(\epsilon^*) \subseteq \text{diag}(\mathcal{E}(\Gamma) \times \mathcal{E}(\Gamma))$ if Γ is weakly acyclic. Combining this property and Theorem 1, the following corollary shows that the action profiles converge to $\mathcal{E}(\Gamma)$ in probability if Γ is weakly acyclic.

Corollary 1 Let Assumptions 1 - 3 hold and Γ is a weakly acyclic game. Consider the sequence of states $\{z(t)\}$ induced by Algorithm 1. Then it holds that $\lim_{t \rightarrow \infty} P\{z(t) \in \text{diag}(\mathcal{E}(\Gamma) \times \mathcal{E}(\Gamma))\} = 1$. Moreover, there exists some $t^v \in \mathbb{N}^0$ such that for any $t^* > t^v$ and $(j-1)(\mathcal{T} + 1) \geq t^*$, inequality (1) holds.

4 Proofs

We will prove Theorem 1 in this section. Let us start with some notations. Fix $t_q, z(t_q) \triangleq (s(t_q), s(t_q + (\mathcal{T} + 1)))$ induced by Algorithm 1 constitutes a time-homogeneous Markov chain with state space \mathcal{Z} and transition matrix $P_q^{\epsilon(t_q)}$. Define stochastic vector $\pi^*(t_q)$ as the stationary distribution of the Markov chain; i.e., $\pi^*(t_q)^T P_q^{\epsilon(t_q)} = \pi^*(t_q)^T$. According to Line 14 in Algorithm 1, when a player performs exploration, it can choose any element in its action set. One can see that, for any pair of states $x, y \in \mathcal{Z}$, y can be reached from x within finite steps. And by Theorem 3.1 in Chapter 6 of Freidlin et al. (2012), $\pi^*(t_q)$ can be represented as follows.

$$\pi^*(t_q) \triangleq \left[\pi_{z_1}^*(\epsilon(t_q)) \cdots \pi_{z_{|\mathcal{Z}|}}^*(\epsilon(t_q)) \right]^T, \quad (2)$$

where

$$\begin{aligned} \pi_{z'}^*(\epsilon(t_q)) &= \frac{\sigma_z(\epsilon(t_q))}{\sum_{z' \in \mathcal{Z}} \sigma_{z'}(\epsilon(t_q))} \\ \sigma_z(\epsilon(t_q)) &= \sum_{T \in G_q(z)} \prod_{(z', z) \in T} P_q^{\epsilon(t_q)}(z', z). \end{aligned}$$

4.1 Stationary distributions without control errors

We first study the properties of the feasible transitions in Markov chains $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_{\mathcal{T}}$ in the following lemma.

Lemma 1 *If Assumptions 1 - 2 hold, then there exists a $t_e > 2\mathcal{T} + 2$ such that for any $q \in \{0, 1, \dots, \mathcal{T}\}$, $j \in \mathbb{N}^0$ and $t_q \geq t_e$, each nonzero entry in transition matrix $P_q^{\epsilon(t_q)}$ of Markov chain \mathcal{M}_q can be characterized as a polynomial of the set variables $\{\epsilon_i(t_q) + e_c^i(t_q), 1 - \epsilon_i(t_q) - e_c^i(t_q)\}$, where $i \in \mathcal{I}$.*

PROOF. Let any two states be $x, y \in \mathcal{Z}$ and $x = (s^0, s^1), y = (s^2, s^3)$, where $s^0 = (a_0^1, \dots, a_0^N), s^1 = (a_1^1, \dots, a_1^N), s^2 = (a_2^1, \dots, a_2^N), s^3 = (a_3^1, \dots, a_3^N)$. Based on the definition of state, if $s^2 \neq s^1$, the transition from x to y is not feasible within one step; i.e., $P_q^{\epsilon(t_q)}(x, y) = 0$ for all $s^2 \neq s^1$. Now we consider a feasible transition from $x = (s^0, s^1)$ to $y = (s^1, s^3)$. s^3 can be achieved by K players performing exploitation and $N - K$ players performing exploration, where $0 \leq K \leq N$. Note that $K = 0$ means all the players perform exploration and $K = N$ means all the players perform exploitation.

According to Line 8 - 13 in Algorithm 1, when a player; e.g., player i , performs exploitation, $a_3^i = a_0^i$ if $\tilde{u}_i(s^0) \geq \tilde{u}_i(s^1)$ and $a_3^i = a_1^i$ if $\tilde{u}_i(s^1) \geq \tilde{u}_i(s^0)$. Without loss of generality, let us consider when $\tilde{u}_i(s^1) \geq \tilde{u}_i(s^0)$. Because of the observation errors, the value of $\tilde{u}_i(s^1)$ could vary from different iterations. Therefore $\tilde{u}_i(s^1) \geq \tilde{u}_i(s^0)$ does not always imply that $u_i(s^1) \geq u_i(s^0)$, and when player i performs exploitation, its choice could be different between a_0^i and a_1^i at different iterations even if $u_i(s^1) \geq u_i(s^0)$. By Assumption 2, for $t_q \geq t_e$, $\tilde{u}_i(s^1) \geq \tilde{u}_i(s^0)$ always implies that $u_i(s^1) \geq u_i(s^0)$, then player i will always choose a_1^i when it performs exploitation. And this holds for any player. When $K = N$, the probability is $\prod_{i=1}^N (1 - \epsilon_i(t_q) - e_c^i(t_q))$ when $t_q \geq t_e$.

According to Line 14 in Algorithm 1, when a player performs exploration, it can choose any element in its action set. Therefore, if an action can be achieved by the player performing exploitation, it can also be achieved by the player performing exploration.

Therefore, when s^3 can be achieved when $K = N$, it actually can also be achieved when $0 \leq K < N$. Now let us consider $0 < K < N$. For any $1 \leq K \leq (N - 1)$, there exist $\binom{N}{K}$ subcases. The difference of the subcases is that different players performing exploration or exploitation. The probabilities of the subcases are still polynomials of the elements in $\{\epsilon_i(t_q) + e_c^i(t_q), 1 - \epsilon_i(t_q) - e_c^i(t_q)\}$, where $i \in \mathcal{I}$. Without loss of generality, let us consider the

first K players perform exploitation and the rest $N - K$ players perform exploration. Then the probability of this subcase is $\prod_{i=1}^K (1 - \epsilon_i(t_q) - e_c^i(t_q)) \times \prod_{i=K+1}^N \frac{(\epsilon_i(t_q) + e_c^i(t_q))}{|\mathcal{A}_i|}$ when $t_q \geq t_e$.

Similarly, when s^3 can be achieved when $0 < K < N$, it actually can also be achieved when $K = 0$ because if an action can be achieved by the player performing exploitation, it can also be achieved by the player performing exploration. The probability when $K = 0$ is $\prod_{i=1}^N \frac{\epsilon_i(t_q) + e_c^i(t_q)}{|\mathcal{A}_i|}$ when $t_q \geq t_e$. Therefore, for $0 \leq K \leq N$, any nonzero entry in transition matrix $P_q^{\epsilon(t_q)}$ must be a linear combination of the set variables $\{\epsilon_i(t_q) + e_c^i(t_q), 1 - \epsilon_i(t_q) - e_c^i(t_q)\}$. It completes the proof of Lemma 1. \square

Now let us consider an auxiliary scenario that control errors $e_c^i(t) = 0$ for all $t = 1, 2, \dots$ and for all $i \in \mathcal{I}$. Fix t_q , denote by $\tilde{P}_q^{\epsilon(t_q)}$ the transition matrix of the time-homogeneous Markov chain constituted by $z(t_q)$. We also define $\tilde{\pi}^*(t_q)$ the stationary distribution of the time-homogeneous Markov chain. It has the same form of (2) with control errors are 0. The following lemma shows that the sequence of stationary distributions $\tilde{\pi}^*(t)$ converges to a limiting distribution with a certain rate, and the support of the limiting distribution is contained in $\Lambda(\epsilon^*)$.

Lemma 2 *If Assumptions 1 - 2 hold and $e_c^i(t) = 0$ for all t and all i , then the sequence of distribution $\{\tilde{\pi}^*(t)\}$, where $t = 0, 1, 2, \dots$, converges to the limiting distribution $\pi^* \triangleq \lim_{t \rightarrow \infty} \tilde{\pi}^*(t)$ whose support is contained in $\Lambda(\epsilon^*)$. Moreover, the convergence rate could be quantified as: $\|\tilde{\pi}^*(t) - \pi^*\| \leq C_V \epsilon_V(t)$ for all $t \geq t_e$, where $\epsilon_V(t) = \max_{i \in \mathcal{I}} \epsilon_i(t)$, where $C_V > 0$ is some constant.*

PROOF. The proof is divided into two claims.

Claim 1. For any $q \in \{0, 1, \dots, \mathcal{T}\}$, the limiting distribution $\lim_{t_q \rightarrow \infty} \tilde{\pi}^*(t_q)$ ¹ exists. And all these limiting distributions are the same; i.e., $\pi^* \triangleq \lim_{t_0 \rightarrow \infty} \tilde{\pi}^*(t_0) = \lim_{t_1 \rightarrow \infty} \tilde{\pi}^*(t_1) = \dots = \lim_{t_{\mathcal{T}} \rightarrow \infty} \tilde{\pi}^*(t_{\mathcal{T}})$. Moreover, the support of π^* is contained in $\Lambda(\epsilon^*)$.

PROOF. Take any $q \in \{0, 1, \dots, \mathcal{T}\}$, by Lemma 1, for all $t_q \geq t_e$ the non-zero entries of $\tilde{P}_q^{\epsilon(t_q)}$ are poly-

¹ Notice that, taking the limit of t_q to infinity means that taking the limit of j to infinity in $q + j(\mathcal{T} + 1)$. And for ease of presentation, $\lim_{j \rightarrow \infty} q + j(\mathcal{T} + 1)$ will be referred to as $\lim_{t_q \rightarrow \infty}$.

nomials of $\{\epsilon_i(t_q), 1 - \epsilon_i(t_q)\}$ since $e_c^i(t_q) = 0$ for all $i \in \mathcal{I}$. Then $\sigma_z(\epsilon(t_q))$ and $\sum_{z' \in \mathcal{Z}} \sigma_{z'}(\epsilon(t_q))$ are polynomials of $\{\epsilon_i(t_q), 1 - \epsilon_i(t_q)\}$. By Assumption 1 - (2), for any $i \neq j$, there exists $\gamma_{ij}, \gamma_{ji} > 0$ such that $\frac{\epsilon_i(t)}{\epsilon_j(t)} = \gamma_{ij}$ and $\frac{\epsilon_j(t)}{\epsilon_i(t)} = \gamma_{ji}$. That is, fix an $i_0 \in \mathcal{I}$, any exploration rate $\epsilon_i(t_q)$ can be represented as $\gamma_{ii_0} \epsilon_{i_0}(t_q)$. Consider $t_q \geq t_e$, then for particular state $z \in \mathcal{Z}$, $\sigma_z(\epsilon(t_q))$ and $\sum_{z' \in \mathcal{Z}} \sigma_{z'}(\epsilon(t_q))$ are polynomials of $\epsilon_{i_0}(t_q)$, and $\tilde{\pi}_z^*(\epsilon(t_q))$ is a ratio of two polynomials of $\epsilon_{i_0}(t_q)$:

$$\tilde{\pi}_z^*(\epsilon(t_q)) = \frac{\alpha_z(\epsilon_{i_0}(t_q))}{\beta(\epsilon_{i_0}(t_q))}. \quad (3)$$

In particular,

$$\begin{aligned} \alpha_z(\epsilon_{i_0}(t_q)) &= b_k^z \epsilon_{i_0}(t_q)^k + b_{k+1}^z \epsilon_{i_0}(t_q)^{k+1} + \dots + b_h^z \epsilon_{i_0}(t_q)^h \\ \beta(\epsilon_{i_0}(t_q)) &= b_k \epsilon_{i_0}(t_q)^k + b_{k+1} \epsilon_{i_0}(t_q)^{k+1} + \dots + b_h \epsilon_{i_0}(t_q)^h, \end{aligned}$$

where $k \geq 0$. Without loss of generality, we assume that of b_k is non-zero.

When $\epsilon_{i_0}(t_q)$ is sufficiently small, $b_k^z \epsilon_{i_0}(t_q)^k$ and $b_k \epsilon_{i_0}(t_q)^k$ dominate $\frac{\alpha_z(\epsilon_{i_0}(t_q))}{\beta(\epsilon_{i_0}(t_q))}$. Then the limit of the $\tilde{\pi}_z^*(\epsilon(t_q))$ can be represented as $\lim_{t_q \rightarrow \infty} \tilde{\pi}_z^*(\epsilon(t_q)) = \frac{b_k^z}{b_k}$.

Given a pair of states (z', z) , when $t_q \geq t_e$, the non-zero transition probability $p_q^{\epsilon(t_q)}(z', z)$ is a polynomial of $\epsilon_{i_0}(t_q)$ and the coefficients are time-homogeneous. Then b_k^z and b_k are also time-homogeneous. That is for any q , $\lim_{t_q \rightarrow \infty} \tilde{\pi}_z^*(\epsilon(t_q))$ are the same. We denote π^* as the limiting distribution.

By Assumption 1 - (1) and $e_c^i(t) = 0$ for all t and all i , we have $\lim_{t \rightarrow \infty} \epsilon(t) = \epsilon^* = 0$, which implies $\lim_{t_q \rightarrow \infty} \epsilon(t_q) = \epsilon^* = 0$ for any q . By the definition of $\Lambda(\epsilon^*)$, if b_k^z is non-zero, then $z \in \Lambda(\epsilon^*)$; And if $b_k^z = 0$, then $z \notin \Lambda(\epsilon^*)$. We know $\pi_z^* = \frac{b_k^z}{b_k}$, therefore the support of π^* is contained in $\Lambda(\epsilon^*)$. It completes the proof of Claim 1. \square

Claim 2. $\|\tilde{\pi}^*(t) - \pi^*\| \leq C_{\vee} \epsilon_{\vee}(t)$ for $t \geq t_e$, where $C_{\vee} > 0$ is some constant.

PROOF. Remind that the sequence $\{t\}$ is partitioned by $\mathcal{T} + 1$ sub-sequences. Therefore for any $t > 0$, there exist a $q \in \{0, 1, \dots, \mathcal{T}\}$ and a $j \in \mathbb{N}^0$ such that $t =$

$q + j(\mathcal{T} + 1)$. Then from Claim 1, we can have $\lim_{t \rightarrow \infty} \tilde{\pi}^*(t) = \pi^*$. Now we consider the convergence rate.

$$\begin{aligned} \|\tilde{\pi}^*(t) - \pi^*\| &= \sum_{z \in \mathcal{Z}} |\tilde{\pi}_z^*(\epsilon(t)) - \pi_z^*| \\ &= \sum_{z \in \Lambda(\epsilon^*)} |\tilde{\pi}_z^*(\epsilon(t)) - \pi_z^*| + \sum_{z \notin \Lambda(\epsilon^*)} |\tilde{\pi}_z^*(\epsilon(t)) - \pi_z^*|. \end{aligned}$$

By Claim 1, we know $\pi_z^* = 0$ when $z \notin \Lambda(\epsilon^*)$ and $\pi_z^* \neq 0$ when $z \in \Lambda(\epsilon^*)$. And note that $\tilde{\pi}_z^*(\epsilon(t))$ is a ratio of two polynomials of $\epsilon_{i_0}(t)$ for $t \geq t_e$.

$$\begin{aligned} \|\tilde{\pi}^*(t) - \pi^*\| &= \sum_{z \in \Lambda(\epsilon^*)} |\tilde{\pi}_z^*(\epsilon(t)) - \pi_z^*| + \sum_{z \notin \Lambda(\epsilon^*)} \tilde{\pi}_z^*(\epsilon(t)) \\ &= \sum_{z \in \Lambda(\epsilon^*)} |\tilde{\pi}_z^*(\epsilon(t)) - \pi_z^*| + 1 - \sum_{z \in \Lambda(\epsilon^*)} \tilde{\pi}_z^*(\epsilon(t)) \\ &= \sum_{z \in \Lambda(\epsilon^*)} |\tilde{\pi}_z^*(\epsilon(t)) - \pi_z^*| + \sum_{z \in \Lambda(\epsilon^*)} \pi_z^* - \sum_{z \in \Lambda(\epsilon^*)} \tilde{\pi}_z^*(\epsilon(t)) \\ &\leq \sum_{z \in \Lambda(\epsilon^*)} |\tilde{\pi}_z^*(\epsilon(t)) - \pi_z^*| + \sum_{z \in \Lambda(\epsilon^*)} |\tilde{\pi}_z^*(\epsilon(t)) - \pi_z^*| \\ &= 2 \sum_{z \in \Lambda(\epsilon^*)} \left| \frac{\alpha_z(\epsilon_{i_0}(t))}{\beta(\epsilon_{i_0}(t))} - \frac{b_k^z}{b_k} \right| \\ &= 2 \sum_{z \in \Lambda(\epsilon^*)} \left| \frac{b_k^z \epsilon_{i_0}(t)^k + \dots + b_h^z \epsilon_{i_0}(t)^h}{b_k \epsilon_{i_0}(t)^k + \dots + b_h \epsilon_{i_0}(t)^h} - \frac{b_k^z}{b_k} \right| \\ &= 2 \sum_{z \in \Lambda(\epsilon^*)} \left| \frac{L_z(\epsilon_{i_0}(t)^{k+1}, \dots, \epsilon_{i_0}(t)^h)}{L(\epsilon_{i_0}(t)^k, \dots, \epsilon_{i_0}(t)^h)} + \frac{\epsilon_{i_0}(t)^k (b_k^z b_k - b_k b_k^z)}{L(\epsilon_{i_0}(t)^k, \dots, \epsilon_{i_0}(t)^h)} \right| \\ &= 2 \epsilon_{i_0}(t) \sum_{z \in \Lambda(\epsilon^*)} \left| \frac{L_z(1, \epsilon_{i_0}(t), \dots, \epsilon_{i_0}(t)^{h-k-1})}{L(1, \epsilon_{i_0}(t), \dots, \epsilon_{i_0}(t)^{h-k})} \right|, \end{aligned}$$

where L_z and L are linear functions, the constant term of L is non-zero. And for any $z \in \Lambda(\epsilon^*)$, $L_z(1, \epsilon_{i_0}(t), \epsilon_{i_0}(t)^{h-k})$ and $L(1, \epsilon_{i_0}(t), \epsilon_{i_0}(t)^{h-k-1})$ converge as $t \rightarrow \infty$. Also because $\Lambda(\epsilon^*)$ contains finite elements, then $2 \sum_{z \in \Lambda(\epsilon^*)} \left| \frac{L_z(1, \epsilon_{i_0}(t), \dots, \epsilon_{i_0}(t)^{h-k-1})}{L(1, \epsilon_{i_0}(t), \dots, \epsilon_{i_0}(t)^{h-k})} \right|$ is uniformly bounded. Remember that we fix i_0 at the beginning. And we can always choose a constant

$C_{\vee} \geq \max_{i_0 \in \mathcal{I}} 2 \sum_{z \in \Lambda(\epsilon^*)} \left| \frac{L_z(1, \epsilon_{i_0}(t), \dots, \epsilon_{i_0}(t)^{h-k-1})}{L(1, \epsilon_{i_0}(t), \dots, \epsilon_{i_0}(t)^{h-k})} \right|$ such that $\|\tilde{\pi}^*(t) - \pi^*\| \leq C_{\vee} \epsilon_{\vee}(t)$, where $\epsilon_{\vee}(t) = \max_{i_0 \in \mathcal{I}} \epsilon_{i_0}(t)$. It completes the proof of Claim 1. \square

It completes the proof of Lemma 2. \square

4.2 Proof of Theorem 1

Lemma 2 shows that $\tilde{\pi}^*(t) \rightarrow \pi^*$ whose support is contained in $\Lambda(\epsilon^*)$. Now we proceed to finish the proofs of Theorem 1 by showing $\pi(t) \rightarrow \pi^*$ and quantifying its convergence rate.

PROOF. We will first prove that for any Markov chain \mathcal{M}_q , its state distribution $\{\pi(t_q)\}$ converges to π^* . Then we prove that the sequence $\{\pi(t)\}$ also converges to π^* . Based on triangle inequality, we can get for any $\pi(q)$ and for all $t_q \geq q + 2\mathcal{T} + 2$,

$$\begin{aligned} \|\pi(t_q) - \pi^*\| &= \left\| \pi(q) P_q^{\epsilon(q+\mathcal{T}+1)} \dots P_q^{\epsilon(t_q)} - \pi^* \right\| \\ &\leq \left\| \pi(q) P_q^{\epsilon(q+\mathcal{T}+1)} \dots P_q^{\epsilon(t_q)} - \pi(q) \tilde{P}_q^{\epsilon(q+\mathcal{T}+1)} \dots \tilde{P}_q^{\epsilon(t_q)} \right\| \\ &\quad + \left\| \pi(q) \tilde{P}_q^{\epsilon(q+\mathcal{T}+1)} \dots \tilde{P}_q^{\epsilon(t_q)} - \pi^* \right\|. \end{aligned} \quad (4)$$

We want to prove that the two terms in the right-hand side of (4) converge to 0 with certain rates.

Claim 3. For any $q \in \{0, 1, \dots, \mathcal{T}\}$,

$\lim_{t_q \rightarrow \infty} \left\| \pi(q) \tilde{P}_q^{\epsilon(q+\mathcal{T}+1)} \dots \tilde{P}_q^{\epsilon(t_q)} - \pi^* \right\| = 0$ and there exists some $t_q^\vee > t_e$ such that for any $t_q^* > t_q^\vee$ and $t_q > t_q^* + \mathcal{T} + 1$, $\left\| \pi(q) \tilde{P}_q^{\epsilon(q+\mathcal{T}+1)} \dots \tilde{P}_q^{\epsilon(t_q)} - \pi^* \right\| \leq C_q \exp\left(-\sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(t) |\mathcal{A}_i|\right) + 4C_\vee \epsilon_\vee(t_q^*) + C_\vee \epsilon_\vee(t_q)$, where C_\vee, C_q are some positive constants.

PROOF. Based on triangle inequality, we can get for all $t_q \geq q + 2\mathcal{T} + 2$,

$$\begin{aligned} \left\| \pi(q) \tilde{P}_q^{\epsilon(q+\mathcal{T}+1)} \dots \tilde{P}_q^{\epsilon(t_q)} - \pi^* \right\| &= \|\tilde{\pi}(t_q) - \pi^*\| \\ &\leq \|\tilde{\pi}(t_q) - \tilde{\pi}^*(t_q)\| + \|\tilde{\pi}^*(t_q) - \pi^*\|, \end{aligned} \quad (5)$$

where $\tilde{\pi}(t_q)$ is the distribution on \mathcal{Z} at t_q when the control errors $e_c^i(t_q) = 0$ for all $i \in \mathcal{I}$. Let $x(t_q) \triangleq \|\tilde{\pi}(t_q) - \tilde{\pi}^*(t_q)\|$ and $y(t_q) \triangleq \|\tilde{\pi}^*(t_q) - \pi^*\|$.

Let us first consider $x(t_q)$. Note that $\tilde{\pi}^*(t_q)^T \tilde{P}_q^{\epsilon(t_q)} = \tilde{\pi}^*(t_q)^T$. Then we have:

$$\begin{aligned} x(t_q) &= \|\tilde{\pi}(t_q) - \tilde{\pi}^*(t_q)\| \\ &= \|\tilde{\pi}(t_q) - \tilde{\pi}^*(t_q - \mathcal{T} - 1) + \tilde{\pi}^*(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q)\| \\ &\leq \left\| \{\tilde{P}_q^{\epsilon(t_q-\mathcal{T}-1)}\}^T \pi(t_q - \mathcal{T} - 1) \right. \\ &\quad \left. - \{\tilde{P}_q^{\epsilon(t_q-\mathcal{T}-1)}\}^T \tilde{\pi}^*(t_q - \mathcal{T} - 1) \right\| \\ &\quad + \|\tilde{\pi}^*(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q)\|. \end{aligned} \quad (6)$$

Based on Lemma 1, when $t_q \geq t_e$, the nonzero entries in $\{\tilde{P}_q^{\epsilon(t_q-\mathcal{T}-1)}\}^T$ can be represented as polynomials of $\{\epsilon_i(t_q - \mathcal{T} - 1), 1 - \epsilon_i(t_q - \mathcal{T} - 1)\}$. Taking the nonzero entry $\prod_{i=1}^N \frac{\epsilon_i(t_q-\mathcal{T}-1)}{|\mathcal{A}_i|}$, we can decompose $\{\tilde{P}_q^{\epsilon(t_q-\mathcal{T}-1)}\}^T$ into the following:

$$\{\tilde{P}_q^{\epsilon(t_q-\mathcal{T}-1)}\}^T = \prod_{i=1}^N \frac{\epsilon_i(t_q - \mathcal{T} - 1)}{|\mathcal{A}_i|} Q + R(t_q - \mathcal{T} - 1),$$

where Q is a $|\mathcal{Z}| \times |\mathcal{Z}|$ matrix with all entries are 1. Because $\tilde{P}_q^{\epsilon(t_q-\mathcal{T}-1)}$ is a transition matrix, then the $\{\tilde{P}_q^{\epsilon(t_q-\mathcal{T}-1)}\}^T$ is a column stochastic matrix where each column sum is equal to 1. It follows that the column sums of $\prod_{i=1}^N \frac{\epsilon_i(t_q-\mathcal{T}-1)}{|\mathcal{A}_i|} Q$ equal

$$\begin{aligned} \prod_{i=1}^N \frac{\epsilon_i(t_q-\mathcal{T}-1)}{|\mathcal{A}_i|} |\mathcal{Z}| &= \prod_{i=1}^N \epsilon_i(t_q - \mathcal{T} - 1) |\mathcal{A}_i| \text{ since} \\ |\mathcal{Z}| &= \left(\prod_{i=1}^N \frac{\epsilon_i(t_q-\mathcal{T}-1)}{|\mathcal{A}_i|} \right)^2, \text{ and the column sums of} \\ R(t_q - \mathcal{T} - 1) &\text{ equal } c(t_q - \mathcal{T} - 1) = 1 - \prod_{i=1}^N \epsilon_i(t_q - \mathcal{T} - 1) |\mathcal{A}_i|. \end{aligned}$$

By (1) in Assumption 1, $\prod_{i=1}^N \epsilon_i(t_q - \mathcal{T} - 1) |\mathcal{A}_i|$ strictly decreases to 0. Then there exists a $t_q^{|\mathcal{A}|}$ such that $\prod_{i=1}^N \epsilon_i(t_q - \mathcal{T} - 1) |\mathcal{A}_i| < 1$ for all $t_q \geq t_q^{|\mathcal{A}|}$, which implies $0 < c(t_q - \mathcal{T} - 1) < 1$ for all $t_q \geq t_q^{|\mathcal{A}|}$ and the column sums of $c(t_q - \mathcal{T} - 1)^{-1} R(t_q - \mathcal{T} - 1)$ equal 1.

Let $v(t_q - \mathcal{T} - 1) \triangleq \|\tilde{\pi}^*(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q)\|$. And consider $t_q \geq t_q^{|\mathcal{A}|}$, then inequality (6) becomes

$$\begin{aligned} x(t_q) &\leq \left\| \{\tilde{P}_q^{\epsilon(t_q-\mathcal{T}-1)}\}^T (\tilde{\pi}(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q - \mathcal{T} - 1)) \right\| \\ &\quad + v(t_q - \mathcal{T} - 1) \\ &= \left\| \prod_{i=1}^N \frac{\epsilon_i(t_q - \mathcal{T} - 1)}{|\mathcal{A}_i|} Q (\tilde{\pi}(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q - \mathcal{T} - 1)) \right. \\ &\quad \left. + R(t_q - \mathcal{T} - 1) (\tilde{\pi}(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q - \mathcal{T} - 1)) \right\| \\ &\quad + v(t_q - \mathcal{T} - 1), \end{aligned} \quad (7)$$

where $\tilde{\pi}(t_q - \mathcal{T} - 1)$ and $\tilde{\pi}^*(t_q - \mathcal{T} - 1)$ are both stochastic vectors whose sum of elements is equal to 1. And by the construction of Q , we have $Q(\tilde{\pi}(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q - \mathcal{T} - 1)) = 0$. Then inequality (7) becomes:

$$\begin{aligned} x(t_q) &\leq c(t_q - \mathcal{T} - 1) \|c(t_q - \mathcal{T} - 1)^{-1} R(t_q - \mathcal{T} - 1) \\ &\quad \times (\tilde{\pi}(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q - \mathcal{T} - 1))\| + v(t_q - \mathcal{T} - 1) \\ &\leq c(t_q - \mathcal{T} - 1) \|c(t_q - \mathcal{T} - 1)^{-1} R(t_q - \mathcal{T} - 1)\| \end{aligned}$$

$$\begin{aligned} & \times ||\tilde{\pi}(t_q - \mathcal{T} - 1) - \tilde{\pi}^*(t_q - \mathcal{T} - 1)|| + v(t_q - \mathcal{T} - 1) \\ & = c(t_q - \mathcal{T} - 1)x(t_q - \mathcal{T} - 1) + v(t_q - \mathcal{T} - 1), \end{aligned}$$

where $c(t_q - \mathcal{T} - 1) \in (0, 1)$ for all $t_q \geq t_q^{|\mathcal{A}|}$.

With inequality $\log(1 - x) < -x, \forall x \in (0, 1)$, for any $t_q > t_q^{|\mathcal{A}|}$ and $t_q^* \geq t_q^{|\mathcal{A}|}$, we have :

$$\begin{aligned} x(t_q) & \leq c(t_q - \mathcal{T} - 1)x(t_q - \mathcal{T} - 1) + v(t_q - \mathcal{T} - 1) \\ & \leq \prod_{\tau_q = t_q^*}^{t_q - \mathcal{T} - 1} c(\tau_q)x(t_q^*) + v(t_q - \mathcal{T} - 1) \\ & \quad + \sum_{\tau_q = t_q^*}^{t_q - 2(\mathcal{T} + 1)} \left(\prod_{i = \tau_q + \mathcal{T} + 1}^{t_q - \mathcal{T} - 1} c(i)v(\tau_q) \right) \\ & \leq x(t_q^*) \prod_{\tau_q = t_q^*}^{t_q - \mathcal{T} - 1} \exp(-(1 - c(\tau_q))) + v(t_q - \mathcal{T} - 1) \\ & \quad + \sum_{\tau_q = t_q^*}^{t_q - 2(\mathcal{T} + 1)} \left(\prod_{i = \tau_q + \mathcal{T} + 1}^{t_q - \mathcal{T} - 1} \exp(-(1 - c(i)))v(\tau_q) \right) \\ & \leq x(t_q^*) \exp\left(-\sum_{\tau_q = t_q^*}^{t_q - \mathcal{T} - 1} (1 - c(\tau_q))\right) + \sum_{\tau_q = t_q^*}^{t_q - \mathcal{T} - 1} v(\tau_q). \end{aligned} \quad (8)$$

In this paper, for any vector, we choose the L^1 -norm, then

$$\begin{aligned} \sum_{\tau_q = q}^{+\infty} v(\tau_q) & = \sum_{\tau_q = q}^{+\infty} ||\tilde{\pi}^*(\tau_q) - \tilde{\pi}^*(\tau_q + \mathcal{T} + 1)|| \\ & = \sum_{\tau_q = q}^{+\infty} \sum_{z \in \mathcal{Z}} |\tilde{\pi}_z^*(\epsilon(\tau_q)) - \tilde{\pi}_z^*(\epsilon(\tau_q + \mathcal{T} + 1))|. \end{aligned} \quad (9)$$

By Assumption 1 - (2), for any $i \neq j$, there exists $\gamma_{ij}, \gamma_{ji} > 0$ such that $\frac{\epsilon_i(t)}{\epsilon_j(t)} = \gamma_{ij}$ and $\frac{\epsilon_j(t)}{\epsilon_i(t)} = \gamma_{ji}$. That is, fix an $i_0 \in \mathcal{I}$, any exploration rate $\epsilon_i(t_q)$ can be represented as $\gamma_{ii_0}\epsilon_{i_0}(t_q)$. We will show $v(\tau_q)$ is summable.

Recall equation (3), for $t_q \geq t_e$ and any $z \in \mathcal{Z}$, $\tilde{\pi}_z^*(\epsilon(t_q)) = \frac{\alpha_z(\epsilon_{i_0}(t_q))}{\beta(\epsilon_{i_0}(t_q))}$ is a ratio of two polynomials of $\epsilon_{i_0}(t_q)$. Then the derivative of $\tilde{\pi}_z^*(\epsilon(t_q))$ is:

$$\begin{aligned} \frac{\partial \tilde{\pi}_z^*(\epsilon(t_q))}{\partial \epsilon_{i_0}(t_q)} & = \frac{1}{\beta(\epsilon_{i_0}(t_q))} \left(\frac{\partial \alpha_z(\epsilon_{i_0}(t_q))}{\partial \epsilon_{i_0}(t_q)} \beta(\epsilon_{i_0}(t_q)) \right. \\ & \quad \left. - \alpha_z(\epsilon_{i_0}(t_q)) \frac{\partial \beta(\epsilon_{i_0}(t_q))}{\partial \epsilon_{i_0}(t_q)} \right), \end{aligned}$$

where $\frac{\partial \alpha_z(\epsilon_{i_0}(t_q))}{\partial \epsilon_{i_0}(t_q)} \beta(\epsilon_{i_0}(t_q)) - \alpha_z(\epsilon_{i_0}(t_q)) \frac{\partial \beta(\epsilon_{i_0}(t_q))}{\partial \epsilon_{i_0}(t_q)}$ is a

polynomial of $\epsilon_{i_0}(t_q)$. So $\frac{\partial \tilde{\pi}_z^*(\epsilon_{i_0}(t_q))}{\partial \epsilon_{i_0}(t_q)}$ can be rewritten as

$$\begin{aligned} & \frac{\partial \tilde{\pi}_z^*(\epsilon(t_q))}{\partial \epsilon_{i_0}(t_q)} \\ & = \frac{c_l^z \epsilon_{i_0}(t_q)^l + c_{l+1}^z \epsilon_{i_0}(t_q)^{l+1} + \dots + c_h^z \epsilon_{i_0}(t_q)^h}{\beta(\epsilon_{i_0}(t_q))}, \end{aligned}$$

where $l \geq 0$, $c_l^z \neq 0$ and $\beta(\epsilon_{i_0}(t_q)) > 0$. When $\epsilon_{i_0}(t_q)$ is sufficiently small, $c_l^z \epsilon_{i_0}(t_q)^l$ dominates the derivative. Therefore, $\exists \hat{\epsilon}_{i_0} > 0$, such that the sign of $\frac{\partial \tilde{\pi}_z^*(\epsilon(t_q))}{\partial \epsilon_{i_0}(t_q)}$ is the sign of c_l^z , $\forall 0 < \epsilon_{i_0}(t_q) \leq \hat{\epsilon}_{i_0}$. By Assumption 1 - (1), $\epsilon_{i_0}(t_q)$ strictly decreases to 0, there exists a \hat{t}_{i_0} such that $\epsilon_{i_0}(\hat{t}_{i_0}) = \hat{\epsilon}_{i_0}$. Therefore, let $t_q^{i_0}$ be smallest iteration index of the Markov chain \mathcal{M}_q such that is larger or equal to $\max\{t_e, \hat{t}_{i_0}\}$. We can define a partition of \mathcal{Z} as follows:

$$\begin{aligned} \mathcal{Z}_1 & = \{z \in \mathcal{Z} | \tilde{\pi}_z^*(\epsilon(t_q)) > \tilde{\pi}_z^*(\epsilon(t_q + \mathcal{T} + 1)), \forall t_q \geq t_q^{i_0}\} \\ \mathcal{Z}_2 & = \{z \in \mathcal{Z} | \tilde{\pi}_z^*(\epsilon(t_q)) < \tilde{\pi}_z^*(\epsilon(t_q + \mathcal{T} + 1)), \forall t_q \geq t_q^{i_0}\}. \end{aligned}$$

Then equality (9) becomes:

$$\begin{aligned} \sum_{\tau_q = q}^{+\infty} v(\tau_q) & = \sum_{\tau_q = q}^{+\infty} \sum_{z \in \mathcal{Z}} |\tilde{\pi}_z^*(\epsilon(\tau_q)) - \tilde{\pi}_z^*(\epsilon(\tau_q + \mathcal{T} + 1))| \\ & \quad + \sum_{\tau_q = t_q^{i_0} + \mathcal{T} + 1}^{+\infty} \sum_{z \in \mathcal{Z}_1} (\tilde{\pi}_z^*(\epsilon(\tau_q)) - \tilde{\pi}_z^*(\epsilon(\tau_q + \mathcal{T} + 1))) \\ & \quad + \sum_{\tau_q = t_q^{i_0} + \mathcal{T} + 1}^{+\infty} \left(1 - \sum_{z \in \mathcal{Z}_1} \tilde{\pi}_z^*(\epsilon(\tau_q + \mathcal{T} + 1)) \right. \\ & \quad \left. - (1 - \sum_{z \in \mathcal{Z}_1} \tilde{\pi}_z^*(\epsilon(\tau_q))) \right) \\ & = \sum_{\tau_q = 0}^{t_q^{i_0}} \sum_{z \in \mathcal{Z}} |\tilde{\pi}_z^*(\epsilon(\tau_q)) - \tilde{\pi}_z^*(\epsilon(\tau_q + \mathcal{T} + 1))| \\ & \quad + 2 \sum_{z \in \mathcal{Z}_1} \tilde{\pi}_z^*(\epsilon(t_q^{i_0} + \mathcal{T} + 1)) - 2 \sum_{z \in \mathcal{Z}_1} \tilde{\pi}_z^*(0) < +\infty. \end{aligned}$$

Note that inequality (8) holds for any $t_q^* \geq t_q^{|\mathcal{A}|}$. And by $v(\tau_q)$ is summable, we first take the limit of t_q and then take the limit of t_q^* , by the summability of $v(\tau_q)$, we can

$$\text{have } \lim_{t_q^* \rightarrow \infty} \lim_{t_q \rightarrow \infty} \sum_{\tau_q = t_q^*}^{t_q - \mathcal{T} - 1} v(\tau_q) = 0.$$

Recall that $1 - c(\tau_q) = \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i|$. If $t_q > t_q^\vee$, where

$t_q^\vee \triangleq \max\{t_q^{i_0}, t_q^{|\mathcal{A}|}\}$, then inequality (8) becomes:

$$\begin{aligned}
x(t) &\leq x(t_q^*) \exp \left(- \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} (1 - c(\tau_q)) \right) + \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} v(\tau_q) \\
&\leq x(t_q^*) \exp \left(- \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i| \right) \\
&\quad + \sum_{\tau_q=t_q^*}^{t_q^\vee} \sum_{z \in \mathcal{Z}} |\tilde{\pi}_z^*(\epsilon(\tau_q)) - \tilde{\pi}_z^*(\epsilon(\tau_q + \mathcal{T} + 1))| \\
&\quad + 2 \sum_{z \in \mathcal{Z}_1} \tilde{\pi}_z^*(\epsilon(t_q^\vee + \mathcal{T} + 1)) - 2 \sum_{z \in \mathcal{Z}_1} \tilde{\pi}_z^*(\epsilon(t_q - \mathcal{T} - 1)).
\end{aligned} \tag{10}$$

Inequality (10) holds for any $t_q^* \geq t_q^{|\mathcal{A}|}$. We take $t_q^* > t_q^\vee$, then (10) becomes:

$$\begin{aligned}
x(t) &\leq x(t_q^*) \exp \left(- \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i| \right) \\
&\quad + 2 \sum_{z \in \mathcal{Z}_1} \tilde{\pi}_z^*(\epsilon(t_q^*)) - 2 \sum_{z \in \mathcal{Z}_1} \tilde{\pi}_z^*(\epsilon(t_q - \mathcal{T} - 1)) \\
&\leq x(t_q^*) \exp \left(- \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i| \right) \\
&\quad + 2 \sum_{z \in \mathcal{Z}} |\tilde{\pi}_z^*(\epsilon(t_q^*)) - \pi_z^* + \pi_z^* - \tilde{\pi}_z^*(\epsilon(t_q - \mathcal{T} - 1))| \\
&\leq x(t_q^*) \exp \left(- \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i| \right) \\
&\quad + 2y(t_q^*) + 2y(t_q - \mathcal{T} - 1).
\end{aligned} \tag{11}$$

Combine inequalities (5) and (11), we can get for any $t_q^* > t_q^\vee$:

$$\begin{aligned}
\|\tilde{\pi}(t_q) - \pi^*\| &\leq x(t_q^*) \exp \left(- \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i| \right) \\
&\quad + 2y(t_q^*) + 2y(t_q - \mathcal{T} - 1) + y(t_q).
\end{aligned} \tag{12}$$

By (3) in Assumption 1, $\prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i|$ is not summable. Therefore, for any $t_q^* > t_q^\vee$, we have

$$\lim_{t_q \rightarrow \infty} x(t_q^*) \exp \left(- \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i| \right) = 0.$$

By Lemma 2, we have

$$\begin{aligned}
&2y(t_q^*) + 2y(t_q - \mathcal{T} - 1) + y(t_q) \\
&\leq 2C_V \epsilon_V(t_q^*) + 2C_V \epsilon_V(t_q - \mathcal{T} - 1) + C_V \epsilon_V(t_q) \\
&\leq 4C_V \epsilon_V(t_q^*) + C_V \epsilon_V(t_q),
\end{aligned}$$

where $2C_V \epsilon_V(t_q^*) + 2C_V \epsilon_V(t_q - \mathcal{T} - 1) \leq 4C_V \epsilon_V(t_q^*)$ since $\epsilon_i(t_q)$ is strictly decreasing to 0. And there exists a positive constant C_q such that $x(t_q^*) \leq C_q$. Therefore, for any $t_q^* > t_q^\vee$ and $t_q > t_q^* + \mathcal{T} + 1$, (12) becomes:

$$\begin{aligned}
\|\tilde{\pi}(t_q) - \pi^*\| &\leq C_q \exp \left(- \sum_{\tau_q=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i| \right) \\
&\quad + 4C_V \epsilon_V(t_q^*) + C_V \epsilon_V(t_q).
\end{aligned}$$

Therefore we reach Claim 3. \square

Claim 4. For any $q \in \{0, 1, \dots, \mathcal{T}\}$, $\lim_{t_q \rightarrow \infty} \|\pi(t_q) - \tilde{\pi}(t_q)\| = 0$ and for $t_q^* > t_e$ and $t_q \geq t_q^* + \mathcal{T} + 1$, $\left\| \pi(q) P_q^{\epsilon(q+\mathcal{T}+1)} \dots P_q^{\epsilon(t_q)} - \pi(q) \tilde{P}_q^{\epsilon(q+\mathcal{T}+1)} \dots \tilde{P}_q^{\epsilon(t_q)} \right\| \leq 2^N (2^N - 1) \prod_{i=1}^N |\mathcal{A}_i| \sum_{k=t_q^*}^{t_q} |e_c^\vee(k)|$.

PROOF. We can get for all $t_q \geq q + 2\mathcal{T} + 2$,

$$\begin{aligned}
&\|\pi(q) P_q^{\epsilon(q+\mathcal{T}+1)} \dots P_q^{\epsilon(t_q)} - \pi(q) \tilde{P}_q^{\epsilon(q+\mathcal{T}+1)} \dots \tilde{P}_q^{\epsilon(t_q)}\| \\
&= \|\pi(t_q) - \tilde{\pi}(t_q)\| = \|\pi(t_q^* - \mathcal{T} - 1) P_q^{\epsilon(t_q^*)} \dots P_q^{\epsilon(t_q)} \\
&\quad - \pi(t_q^* - \mathcal{T} - 1) \tilde{P}_q^{\epsilon(t_q^*)} \dots \tilde{P}_q^{\epsilon(t_q)}\| \\
&\leq \sum_{k=t_q^*}^{t_q} \|P_q^{\epsilon(k)} - \tilde{P}_q^{\epsilon(k)}\| \\
&\leq \sum_{x, y \in \mathcal{Z}} \sum_{k=t_q^*}^{t_q} \left| P_q^{\epsilon(k)}(x, y) - \tilde{P}_q^{\epsilon(k)}(x, y) \right|.
\end{aligned} \tag{13}$$

The first inequality of (13) holds because $\|\pi(t_q^* - \mathcal{T} - 1) P_q^{\epsilon(t_q^*)} \dots P_q^{\epsilon(t_q)} - \pi(t_q^* - \mathcal{T} - 1) \tilde{P}_q^{\epsilon(t_q^*)} \dots \tilde{P}_q^{\epsilon(t_q)}\| \leq \|\tilde{P}_q^{\epsilon(t_q^*)} \dots P_q^{\epsilon(t_q)} - \tilde{P}_q^{\epsilon(t_q^*)} \dots \tilde{P}_q^{\epsilon(t_q)}\|$ and $\|\tilde{P}_q^{\epsilon(t_q^*)} \dots P_q^{\epsilon(t_q)} - \tilde{P}_q^{\epsilon(t_q^*)} \dots \tilde{P}_q^{\epsilon(t_q)}\| \leq \sum_{k=t_q^*}^{t_q} \|P_q^{\epsilon(k)} - \tilde{P}_q^{\epsilon(k)}\|$ (proved in (7) of Gutjahr and Pflug (1996)).

By Lemma 1, for any $k \geq t_e$, any entry in $P_q^{\epsilon(k)}$ can be represented as a summation of at most $\sum_{j=0}^N \binom{N}{j} = 2^N$

polynomials (j players perform exploration and $N - j$ players perform exploration). And each polynomial is a product of N binomials; e.g., $\prod_{i=1}^N \frac{\epsilon_i(k) + e_c^i(k)}{|\mathcal{A}_i|}$.

And the entries in $\tilde{P}_q^{\epsilon(k)}$ have the same form with $e_c^i(k) = 0$. Then the difference of any pair of entries $(P_q^{\epsilon(k)}(x, y), \tilde{P}_q^{\epsilon(k)}(x, y))$ has at most $2^N(2^N - 1)$ terms (for example, $\prod_{i=1}^N \frac{\epsilon_i(k) + e_c^i(k)}{|\mathcal{A}_i|} - \prod_{i=1}^N \frac{\epsilon_i(t)}{|\mathcal{A}_i|}$ has $2^N - 1$ terms). And each term is less than $\prod_{i=1}^N \frac{|e_c^\vee(k)|}{|\mathcal{A}_i|}$, where $e_c^\vee(t) = \max\{e_c^1(t), \dots, e_c^N(t)\}$. Then any pair of entries $(P_q^{\epsilon(k)}(x, y), \tilde{P}_q^{\epsilon(k)}(x, y))$ of transition matrix $P_q^{\epsilon(k)}$ and transition matrix $\tilde{P}_q^{\epsilon(k)}$ satisfy that:

$$|P_q^{\epsilon(k)}(x, y) - \tilde{P}_q^{\epsilon(k)}(x, y)| \leq 2^N(2^N - 1) \prod_{i=1}^N \frac{1}{|\mathcal{A}_i|} |e_c^\vee(k)|. \quad (14)$$

By Assumption 3, we have $2^N(2^N - 1) \prod_{i=1}^N \frac{1}{|\mathcal{A}_i|} \sum_{k=t_q^*}^{\infty} |e_c^\vee(k)| < \infty$. Therefore, $\|\pi(0)P_q^{\epsilon(1)} \dots P_q^{\epsilon(t_q)} - \pi(0)\tilde{P}_q^{\epsilon(1)} \dots \tilde{P}_q^{\epsilon(t_q)}\| \leq \sum_{x, y \in \mathcal{Z}} \sum_{k=t_q^*}^{t_q} |P_q^{\epsilon(k)}(x, y) - \tilde{P}_q^{\epsilon(k)}(x, y)| \leq 2^N(2^N - 1) \prod_{i=1}^N |\mathcal{A}_i| \sum_{k=t_q^*}^{t_q} |e_c^\vee(k)|$.

We first take the limit of t_q and then take the limit of t_q^* , by the summability of $|e_c(k)|$, we can have

$$\lim_{t_q^* \rightarrow \infty} \lim_{t_q \rightarrow \infty} 2^N(2^N - 1) \prod_{i=1}^N |\mathcal{A}_i| \sum_{k=t_q^*}^{t_q} |e_c^\vee(k)| = 0. \text{ Therefore we reach Claim 4. } \square$$

Combining Claim 3 and Claim 4, we get that for any Markov chain \mathcal{M}_q , its state distribution $\{\pi(t_q)\}$ converges to same limiting distribution π^* .

Now we consider the sequence of $\{\pi(t)\}$. For any $q \in \{0, 1, \dots, \mathcal{T}\}$, and any $\varepsilon > 0$, there exists an integer $\mathcal{N}(q)$ such that $\|\pi(t_q) - \pi^*\| < \varepsilon$. And for any $t > 0$, we can find a $q \in \{0, 1, \dots, \mathcal{T}\}$ and a $j \in \mathbb{N}^0$ such that t can be represented as $t = t_q + j(\mathcal{T} + 1)$. Now we take $t > \max_q \mathcal{N}(q)$, then $\|\pi(t) - \pi^*\| < \varepsilon$. Therefore, the sequence $\{\pi(t)\}$ converges to π^* as well.

Moreover, by triangle inequality, Claim 3 and Claim 4, for any $q \in \{0, 1, \dots, \mathcal{T}\}$, and its corresponding $t_q^\vee > t_e$

such that for any $t_q^* > t_q^\vee$ and $t_q > t_q^* + \mathcal{T} + 1$,

$$\begin{aligned} \|\pi(t_q) - \pi^*\| &\leq C_q \exp\left(-\sum_{\tau=t_q^*}^{t_q-\mathcal{T}-1} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i|\right) \\ &\quad + 4C_\vee \epsilon_\vee(t_q^*) \\ &\quad + C_\vee \epsilon_\vee(t_q) + 2^N(2^N - 1) \prod_{i=1}^N |\mathcal{A}_i| \sum_{k=t_q^*}^t |e_c^\vee(k)|. \end{aligned}$$

Then there exists some $t^\vee = \max_q \{t_q^\vee\}$, $t^* = \min_q \{t_q^*\}$ and $C_e = \max_q \{C_q\}$, for any j such that for any $t^* > t^\vee$ and $(j - 1)(\mathcal{T} + 1) \geq t^*$, we have,

$$\begin{aligned} D(j) &= \sum_{q=0}^{\mathcal{T}} \|\pi(q + j(\mathcal{T} + 1)) - \pi^*\| \\ &\leq (\mathcal{T} + 1) \max_q \{\|\pi(t_q) - \pi^*\|\} \\ &\leq (\mathcal{T} + 1) \left(C_e \exp\left(-\sum_{\tau=t^*}^{(j-1)(\mathcal{T}+1)} \prod_{i=1}^N \epsilon_i(\tau_q) |\mathcal{A}_i|\right) \right. \\ &\quad \left. + 4C_\vee \epsilon_\vee(t^*) + C_\vee \epsilon_\vee(j(\mathcal{T} + 1)) \right. \\ &\quad \left. + 2^N(2^N - 1) \prod_{i=1}^N |\mathcal{A}_i| \sum_{\tau=t^*}^t |e_c^\vee(\tau)| \right). \end{aligned}$$

It completes the proof of Theorem 1. \square

5 Case studies

In this section, we will apply Algorithm 1 to two applications; i.e., the demand allocation market in Zhu (2014) and the cyber security scenario in Okhravi et al. (2014), to demonstrate the performance of Algorithm 1.

5.1 Case 1: demand allocation market

In this section, we study a power market which consists of N customers and a system operator. Each customer wants to allocate its demands in the near future time slots and its action is subject to the prices enforced by the system operator.

5.1.1 System components

Customers. We consider N customers $\mathcal{I} = \{1, \dots, N\}$ and each customer $i \in \mathcal{I}$ has power demands $x_i \geq 0$ and wants to allocate its demands in one time slot within $\mathcal{A}_i = \{1, 2, \dots, |\mathcal{A}_i|\}$. The action $a^i \in \mathcal{A}_i$ is the time slot chosen by customer i . Each customer wants to satisfy its demands as soon as possible so it punishes the late allocation. The cost function $c_i : \mathcal{A}_i \rightarrow \mathbb{R}$ is not decreasing; i.e., $c_i(a^i) \leq c_i(\hat{a}^i)$ if $\hat{a}^i > a^i$.

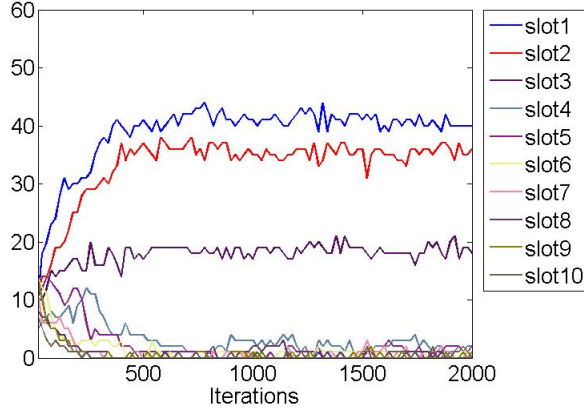


Fig. 3. Temporal aggregate demand allocations

System operator. The system operator charges each customer some price based on demand distributions. In particular, given an action profile $s = (a^1, \dots, a^N)$, the total demand allocated in time slot a^i is $\Xi_{a^i}(s) \triangleq \sum_{j \in \mathcal{I}} \mathbf{1}_{\{a^j = a^i\}} x_j$, where $\mathbf{1}_{\{\Pi\}}$ is an indicator function: $\mathbf{1}_{\{\Pi\}} = 1$ if Π is true and $\mathbf{1}_{\{\Pi\}} = 0$ if Π is false. The system operator charges customer i the price $p_a(\Xi_{a^i}(s))$.

Utility. The utility of customer i is the negative of the cost and price: $u_i(s) = -c_i(a^i) - p_a(\Xi_{a^i}(s))$.

Information constraint. Each customer is unwilling to share its cost function c_i and private action a^i with other customers and the system operator. And the system operator does not want to disclose the pricing policy to the customers and only agrees to publicize the price value $p_a(s)$ given s . Therefore, each customer only knows its own utility values instead of the structure of the utility function.

5.1.2 Evaluation

Evaluation setup. In this section, we use Matlab simulations to evaluate the performance of Algorithm 1. Similar to the setup in Zhu (2014), we consider 100 customers and they have identical action sets consisting of 10 time slots. The demands of all customers are 1; i.e., $x_i = 1$ for any $i \in \mathcal{I}$. The cost function for customer i is set as $c_i(a^i) = \rho_i \xi_i^{a^i}$, where $\rho_i > 0$ and $\xi_i > 1$. And the pricing mechanism is $p_a(\Xi_{a^i}(s)) = \Xi_{a^i}(s)$.

Nash equilibrium. By Lemma 2.1 in Zhu (2014), we know that the demand allocation game under the above setup is a potential game, and then a weakly acyclic game Monderer and Shapley (1996). Therefore the existence of pure Nash equilibrium is guaranteed.

Matlab simulation results. Based on the evaluation setup, we simulate the interactions of the customers and

system operator in Matlab. Where the exploration rates are chosen as $\epsilon_i(t) = \frac{1}{10} t^{-\frac{1}{100}}$ and the control errors are chosen as $e_c^i(t) = \frac{1}{10^{100} t^2}$ for all $i \in \mathcal{I}$. The utility errors are chosen as $e_o^i(t) = \frac{1}{t^2}$ and the upper bound of utility delays is considered as 1 iteration for all $i \in \mathcal{I}$. The duration of the simulation is 2,000 iterations. In Figure 3, each curve represents the temporal aggregate demands allocated at the particular time slot. Figure 3 suggests the convergence of the action profiles.

5.2 Case 2: cyber security scenario

In this section, we study a real-world cyber security scenario which consists of two players: the defender and attacker. The system is the server containing several zero-day security vulnerabilities. A zero-day attack happens once that a software/hardware vulnerability is exploited by the attacker before software the engineers develop any patch to fix the vulnerability. The attacker is equipped with a set of zero-day attack scripts denoted as \mathcal{A} and the defender is equipped with a set of platforms denoted as \mathcal{D} . The defender uses a defensive technique called dynamic platforms Okhravi et al. (2014), which changes the properties of the server such that it is harder for the attacker to succeed. The components in the cyber security scenario and the interactions among the components will be discussed in the following paragraphs.

5.2.1 System components

Defender. The defender has a set of different platforms; e.g., different versions of operating systems and architectures. The defender periodically restarts the server, chooses one platform from \mathcal{D} and deploys it on the server each time it restarts the server. The iteration denoted in Section 2.1 is the defense period in this scenario. The action $d(t)$ is the platform deployed at iteration t .

Attacker. The attacker has a set of zero-day attack scripts, where each attack script can only succeed on some platforms, but no the others. The attacker periodically chooses one of the attack scripts to attack the server. Notice that the attack period is often smaller than the defense period because the defender cannot restart the server too frequently due to the resource consumption of restarting the server. In fact, the defense period is usually a multiple of the attack period. The attack action at iteration t , denoted as $a(t)$, is a subset of the attack scripts and the attack action set \mathcal{A} includes all possible subsets. The order of choosing the attack scripts in one iteration does not matter.

Server. Once an attack action succeeds, the attacker can control the server for a certain amount of time. And every time the server restarts, the defender takes over the control of the server. And here we assume the time consumed by the attack scripts to succeed and the time

Table 1. Utility table

Defense actions	Attack actions										
	(0,10)	(1,9)	(2,8)	(3,7)	(4,6)	(5,5)	(6,4)	(7,3)	(8,2)	(9,1)	(10,0)
d_1 (Fedora 11)	0,1.0	0.1,0.9	0.2,0.8	0.3,0.7	0.4,0.6	0.5,0.5	0.6,0.4	0.7,0.3	0.8,0.2	0.9,0.1	1.0,0
d_2 (Gentoo 9)	1.0,0	0.9,0.1	0.8,0.2	0.7,0.3	0.6,0.4	0.5,0.5	0.4,0.6	0.3,0.7	0.2,0.8	0.1,0.9	0,1.0
d_3 (CentOS 6.3)	0,1.0	0.1,0.9	0.2,0.8	0.3,0.7	0.4,0.6	0.5,0.5	0.6,0.4	0.7,0.3	0.8,0.2	0.9,0.1	1.0,0
d_4 (Debian 6)	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0
d_5 (FreeBSD 9)	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0	1.0,0

consumed by restarting the server are negligible compared with the length of an iteration.

Utility. The goal of both the attacker and defender is to gain longer control time of the server. The utility of the defender $u_d(d(t), a(t))$ is the fraction of the time controlled by the defender during iteration t and the utility of the attacker $u_a(d(t), a(t))$ is the fraction of the time controlled by the attacker. Notice that $u_d(d(t), a(t)) + u_a(d(t), a(t)) = 1$.

Information constraint. The attacker can observe when the server restarts, so it knows the iterations, but it does not know the defender's action set \mathcal{D} and which platform is deployed. The defender does not know the attacker's action set \mathcal{A} and the specific attack scripts chosen by the attacker. At the end of each defense period, both the defender and attacker can measure how much time they control the server. Therefore, each player only knows its own utility values instead of the structure of the utility function.

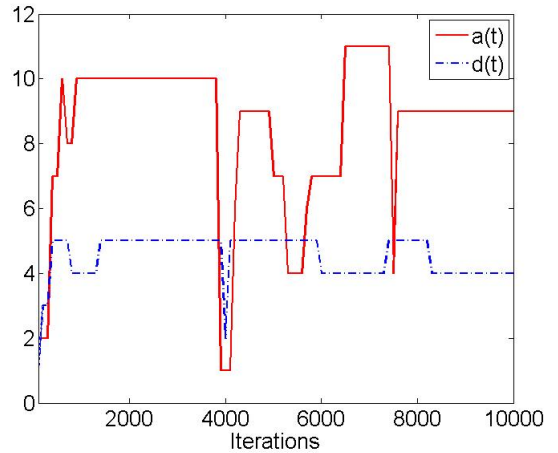
5.2.2 Evaluation

Evaluation setup. In this section, we use Matlab simulations to evaluate the performance of our algorithm based on real-world platform settings, attack scripts and server control data (Okhravi et al., 2012, 2014). The total number of defense actions is 5; i.e., the defender has five different platforms: Fedora 11 on x86, Gentoo 9 on x86, Debian 6 on x86, FreeBSD 9 on x86, and CentOS 6.3 on x86. The attacker has two zero-day attack scripts: TCP MAXSEG exploit, and Socket Pairs exploit. The defense period is set to be ten times as large as the attack period; i.e., during one iteration, the attacker launches 10 attack scripts. Since the time consumed by the attack scripts to succeed is negligible, one attack script enables the attacker control $\frac{1}{10}$ of the iteration if it succeeds. The total number of attack actions is 11; i.e., $a_1 = (0, 10), a_2 = (1, 9), \dots, a_{11} = (10, 0)$, where $a(t) = (0, 10)$ means the attacker launches 0 TCP MAXSEG exploit and 10 Socket Pairs exploits at iteration t .

Real-world utility values. Based on the evaluation setup and the real-world attack scripts, we first replay different attack actions on different platforms to get the

utility table for the defender and the attacker. The results are shown in Table 1, where the defender is the row player and the attacker is the column player. In each cell, the first number represents the utility value to the defender, and the second number represents the utility value to the attacker.

Nash equilibrium. By Proposition 1 in Takahashi and Yamamori (2002), we know any 2-player finite game and its any sub-game (any game constructed by restricting the set of actions to a subset of the set of actions in the original game) has at least one pure Nash equilibrium is a weakly acyclic game. From Table 1, we can see any sub-game has at least one pure Nash equilibrium. Now we want to calculate the pure Nash equilibrium (equilibria). From Table 1, we can see if the defense strategy is d_4 (deploying Debian 6) or d_5 (deploying FreeBSD 9), then the utility of the defender is 1 (the utility of the attacker is 0) not matter what action the attacker uses. From Definition 1, we know the combinations of any attacker action and defense action d_4 or d_5 are pure Nash equilibria.

Fig. 4. Trajectories of $a(t)$ and $d(t)$

Matlab simulation results. Based on Table 1, we simulate the interactions of the defender and attacker in Matlab. Where the exploration rates are chosen as $\epsilon_d(t) = \frac{1}{\sqrt{55t}}, \epsilon_a(t) = \frac{1}{\sqrt{55t}}$ and the control errors are

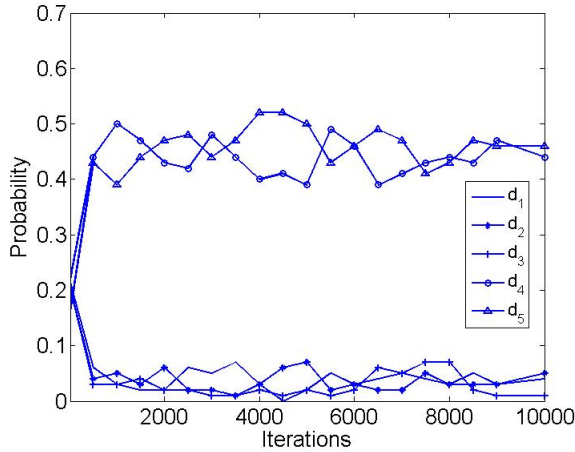


Fig. 5. The probabilities trends of choosing defense actions

chosen as $e_c^d(t) = \frac{1}{110t^2}$, $e_c^a(t) = \frac{1}{110t^2}$. The upper bound of utility delays is considered as 1 iteration in the simulation. The duration of each simulation (from the attack begins till the attack ends) is 10,000 iterations and we repeat 100 identical simulations. Figure 4 shows the trajectories of the defense and attack actions in one certain simulation. And for each simulation, we record the defense action at each iteration. Then at each iteration t , we have 100 chosen defense actions and we use the number of each defense action over 100 as the probability of choosing such defense action at t . Figure 5 shows that the probabilities of choosing different defense actions at different iterations. The result in Figure 4 suggests that the defense action converges to the set $\{d_4, d_5\}$. And the result in Figure 5 shows that the probabilities of choosing defense action d_4 or d_5 are much higher than other 3 actions. Besides, the probabilities of choosing d_4 or d_5 get higher as time goes on. Notice that the combinations of any attacker action and defense action d_4 or d_5 are pure Nash equilibria. Then the simulation results confirm that the convergence of the action profiles to the set of pure Nash equilibria.

6 Conclusion

This paper investigates a class of multi-player discrete games where each player aims to maximize its own utility function with limited information about the game of interest. We propose a robust adaptive learning algorithm which converges in probability to the set of action profiles which have maximal stochastic potential. The convergence rate of the proposed algorithm is analytically quantified. Moreover, the performance of the algorithm is verified by two case studies.

References

Altman, E., Basar, T., Srikant, R., Jun 2002. Nash equilibria for combined flow control and routing in

networks: asymptotic behavior for a large number of users. *IEEE Transactions on Automatic Control* 47 (6), 917–930.

- Arrow, K., Debreu, G., 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22, 265–290.
- Arslan, G., Marden, J. R., Shamma, J. S., 2007. Autonomous vehicle-target assignment: A game-theoretical formulation. *ASME Journal on Dynamic Systems, Measurement, and Control* 129 (5), 584–596.
- Basar, T., Olsder, G., 1999. Dynamic noncooperative game theory. *SIAM Classics in Applied Mathematics*.
- Fabrikant, A., Jaggar, A. D., Schapira, M., Oct 2010. On the structure of weakly acyclic games. In: *International Symposium on Algorithmic Game Theory (SAGT 2010)*. Athens, Greece, pp. 126–137.
- Facchinei, F., Kanzow, C., 2007. Generalized Nash equilibrium problems. *4OR* 5 (3), 173–210.
- Foster, D., Young, P., 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38 (2), 219 – 232.
- Freidlin, M. I., Szücs, J., Wentzell, A. D., 2012. Random perturbations of dynamical systems. Vol. 260. Springer Science & Business Media.
- Frihauf, P., Krstic, M., Basar, T., 2012. Nash equilibrium seeking in non-cooperative games. *IEEE Transactions on Automatic Control* 57 (5), 1192–1207.
- Fudenberg, D., Levine, D. K., 1998. The theory of learning in games. Vol. 2. MIT press.
- Gutjahr, W. J., Pflug, G. C., 1996. Simulated annealing for noisy cost functions. *Journal of Global Optimization* 8 (1), 1–13.
- Hatanaka, T., Wasa, Y., Funada, R., Charalambides, A. G., Fujita, M., 2016. A payoff-based learning approach to cooperative environmental monitoring for ptz visual sensor networks. *IEEE Transactions on Automatic Control* 61 (3), 709–724.
- Liu, S., Krstic, M., 2011. Stochastic Nash equilibrium seeking for games with general nonlinear payoffs. *SIAM Journal on Control and Optimization* 49 (4), 1659–1679.
- Marden, J., Young, H., Arslan, G., Shamma, J., February 2009. Payoff based dynamics for multi-player weakly acyclic games. *SIAM Journal on Control and Optimization* 48 (1), 373–396.
- Marden, J. R., Ruben, S. D., Pao, L. Y., July 2013. A model-free approach to wind farm control using game theoretic methods. *IEEE Transactions on Control Systems Technology* 21 (4), 1207–1214.
- Milchtaich, I., 1996. Congestion games with player-specific payoff functions. *Games and economic behavior* 13 (1), 111–124.
- Monderer, D., Shapley, L., 1996. Potential games. *Games and Economic Behavior* 14 (1), 124 – 143.
- Okhravi, H., Comella, A., Robinson, E., Haines, J., 2012. Creating a cyber moving target for critical infrastructure applications using platform diversity. *International Journal of Critical Infrastructure Protection* 5 (1), 30–39.
- Okhravi, H., Riordan, J., Carter, K., Sep 2014. Quanti-

- tative evaluation of dynamic platform techniques as a defensive mechanism. In: Research in Attacks, Intrusions and Defenses: 17th International Symposium, RAID 2014. Gothenburg, Sweden, pp. 405–425.
- Palomar, D., Eldar, Y., 2010. Convex optimization in signal processing and communications. Cambridge University Press.
- Pang, J.-S., Scutari, G., Facchinei, F., Wang, C., 2008. Distributed power allocation with rate constraints in Gaussian parallel interference channels. *IEEE Transactions on Information Theory* 54 (8), 3471–3489.
- Rosen, J., 1965. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica* 33 (3), 520–534.
- Rouboutsos, A., Kapros, S., 2008. A game theory approach to urban public transport integration policy. *Transport Policy* 15 (4), 209 – 215.
- Sandholm, W. H., 2010. Population games and evolutionary dynamics. MIT press.
- Stankovic, M., Johansson, K., Stipanovic, D., 2012. Distributed seeking of Nash equilibria with applications to mobile sensor networks. *IEEE Transactions on Automatic Control* 57 (4), 904–919.
- Takahashi, S., Yamamori, T., 2002. The pure Nash equilibrium property and the quasi-acyclic condition. *Economics bulletin* 3 (22), 1–6.
- Wang, G., Shanbhag, U. V., Meyn, S. P., Dec 2012. On Nash equilibria in duopolistic power markets subject to make-whole uplift. In: 2012 IEEE 51st IEEE Conference on Decision and Control (CDC). Maui, Hawaii, USA, pp. 472–477.
- Yin, H., Shanbhag, U., Mehta, P., 2011. Nash equilibrium problems with scaled congestion costs and shared constraints. *IEEE Transactions on Automatic Control* 56 (7), 1702–1708.
- Young, H. P., 1993. The evolution of conventions. *Econometrica: Journal of the Econometric Society* 61, 57–84.
- Young, H. P., 2001. Individual strategy and social structure: An evolutionary theory of institutions. Princeton University Press.
- Zhu, M., July 2014. Distributed demand response algorithms against semi-honest adversaries. In: IEEE PES General Meeting. National Harbor, MD, no. 943.
- Zhu, M., Frazzoli, E., 2016. Distributed robust adaptive equilibrium computation for generalized convex games. *Automatica* 63 (1), 82–91.
- Zhu, M., Hu, Z., Liu, P., Nov 2014. Reinforcement learning algorithms for adaptive cyber defense against Heartbleed. In: First ACM Workshop on Moving Target Defense (MTD ’14). Scottsdale, Arizona, USA, pp. 51–58.
- Zhu, M., Martinez, S., 2013. Distributed coverage games for energy-aware mobile sensor networks. *SIAM Journal on Control and Optimization* 51 (1), 1–27.